

Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online

Rimbun Siringoringo^{1*}, Jamaluddin²

Address : Universitas Methodist Indonesia, Manajemen Informatika, Indonesia^{1,2}

Email : rimbun.ringo@gmail.com¹, jac.satuno@gmail.com²

* Corresponding author

Abstrak

Pertumbuhan media sosial dan *e-commerce* mengubah cara berinteraksi dan menyampaikan pandangan, opini dan *mood*. Ulasan produk merupakan salah satu bentuk penyampaian opini dan sentimen konsumen terhadap sebuah produk secara *online*. Ulasan produk saat ini memiliki peranan yang sangat penting dalam mempengaruhi minat konsumen terhadap sebuah produk. Analisis sentimen merupakan pendekatan yang banyak dikerjakan untuk mengekstrak informasi dan menggali opini berkaitan dengan ulasan produk. Analisis sentimen memiliki beberapa tantangan, yang pertama sering sekali hasil analisis sentimen yang dihasilkan oleh model-model prediksi berbeda dengan sentimen yang aktual, tantangan kedua adalah berkaitan dengan cara konsumen mengekspresikan sentimen dan *mood* selalu berbeda dari satu keadaan ke keadaan berikutnya. Pada penelitian ini dilakukan analisis sentimen berdasarkan ulasan produk sepatu *Trendy Shoes* merek Denim. Tahapan analisis sentimen terdiri dari pengumpulan data, pemrosesan awal, transformasi data, seleksi fitur dan tahapan klasifikasi menggunakan *Support Vector Machine*. Pemrosesan awal menerapkan tahapan *text mining* yakni *case folding*, *non alpha numeric removal*, *stop words removal*, dan *stemming*. Hasil analisis sentimen diukur menggunakan kriteria *Akurasi*, *G-Mean*, dan *F-Measure*. Dengan menerapkan pengujian pada tiga jenis data sentimen diperoleh hasil bahwa *Support Vector Machine* dapat mengklasifikasi sentimen dengan baik. Performa *Support Vector Machine* dibandingkan dengan metode *K-Nearest Neighbor*. Hasil klasifikasi sentimen menggunakan *Support Vector Machine* lebih unggul dari *K-Nearest Neighbor*.

Keywords – *sentiment analysis, text mining, support vector machine, product review*

1. Pendahuluan

Hasil riset yang diadakan oleh lembaga riset *Marketing Research* (Nielsen, 2014), mengemukakan fakta bahwa sebanyak 71 % masyarakat pengguna internet saat ini melakukan survei sebelum melakukan aksi jual-beli secara online. Ulasan produk memiliki pengaruh yang sangat besar dalam mempengaruhi minat konsumen. Berdasarkan hasil riset dari *Dimensional Research*, terdapat 91 % calon pembeli memutuskan untuk membeli produk karena dikuatkan oleh ulasan produk positif, sebesar 86 % calon pembeli memutuskan untuk tidak membeli produk tertentu karena dikuatkan oleh ulasan produk negatif [1].

Analisis sentimen merupakan studi dalam rangka menganalisis opini, emosi, dan sentimen masyarakat terkait suatu hal [2]. Analisis Sentimen menerapkan *Natural Language Processing* (NLP) dan *text mining* dalam mengidentifikasi dan mengekstrak informasi tentang topik tertentu [3]. Mengingat

perkebembangan media sosial, *e-commerce*, dan blog, saat ini analisis sentimen penting bagi kehidupan sosial-ekonomi masyarakat.

Ada dua tantangan utama dalam melakukan analisis sentimen, pertama sering sekali bahwa sentimen yang dihasilkan oleh model-model prediksi berbeda dengan sentimen yang aktual, kedua cara konsumen mengekspresikan emosi dan sentimen selalu berbeda beda dari waktu ke waktu [4]

Data mining dan *text mining* menawarkan pendekatan komputasional dalam menganalisis sentimen. *Data mining* juga telah menjadi pendekatan yang paling populer saat ini untuk menggali opini publik [5]. Beberapa penelitian terkait penerapan *data mining* pada sentimen analisis adalah penerapan *Naive Bayes Classifier* [2] pada analisis sentimen terhadap produk buku dan elektronik berdasarkan ulasan produk yang dikumpulkan dari situs jual beli *Amazon.com*. Penerapan metode *K-Means Clustering* (KMC) dan

Decision Tree pada analisis sentimen untuk mengidentifikasi opini konsumen terhadap produk baru [6]. Metode *Support Vector Machine* (SVM) diterapkan pada analisis sentimen terhadap *film box-office* berdasarkan ulasan produk pada situs *imdb.com* [4]

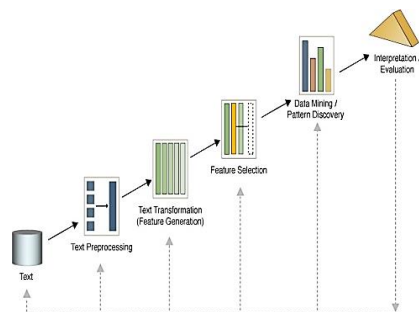
2. Studi Literatur

2.1 Analisis Sentimen

Salah satu gol dari analisis sentimen adalah mengidentifikasi, menggali, serta mengekstrak opini publik terkait dengan topik tertentu [7], sehingga analisis sentimen sering juga dikenal dengan istilah *opinion mining* [2]. Salah satu masalah pada analisis sentimen adalah bagaimana menentukan kategori sentimen pada teks, apakah positif atau negatif.

2.2 Text Mining

Analisis sentimen dan *text mining* merupakan dua hal yang tidak dapat dipisahkan terutama jika melakukan analisis sentimen pada media *online*. *Text mining* merupakan studi yang lebih spesifik pada *data mining* untuk mengungkap pola yang tersembunyi pada teks, sehingga ketika analisis sentimen dan *text mining* dipadukan pada penggalian opini, akan dihasilkan alat bantu yang sangat baik dan andal [7].



Gambar 1. Alur *text mining*

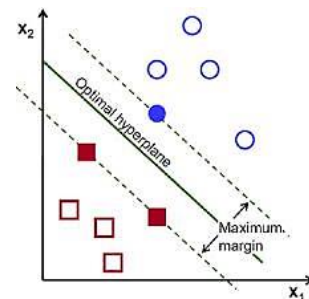
Berdasarkan gambar 1, alur *text mining* pada umumnya adalah tahap pengumpulan teks, pemrosesan awal atau *text pre-processing*, transformasi teks atau *text trasformation*, seleksi fitur atau *feature selection*, *data mining* dan interpretasi [8] Keseluruhan tahapan tersebut akan menjadi acuan pada penelitian ini.

Pemrosesan awal data atau *text pre-processing* terdiri dari *case folding*, *non alpha numeric removal*, *stop words removal*, dan *stemming*. Tahap *case folding* bertujuan untuk mengubah bentuk teks kedalam bentuk huruf kecil, *non alpha numeric removal* bertujuan untuk menghilangkan semua karakter selain karakter alfanumerik, seperti !, @, &, dan yang lainnya. *Stop words* adalah kata-kata yang bukan merupakan ciri (kata unik) dari suatu dokumen seperti

kata sambung atau kata kepunyaan, *stop words removal* bertujuan untuk menghilangkan daftar *stop words* dari dokumen. *Stemming* bertujuan untuk mengembalikan sebuah kata ke dalam bentuk dasarnya, misalnya kata *mempermainkan* menjadi *main*, *peranan* menjadi *peran*. Selanjutnya dilakukan transformasi data untuk mengubah data string menjadi numerik agar dapat diproses oleh algoritma *machine learning*

2.2 Support Vector Machine

Secara sederhana konsep Support Vector Machine (SVM) dapat dijelaskan sebagai usaha mencari *hyperplane-hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada input ruang vektor



Gambar 2 Hyperlane pada SVM

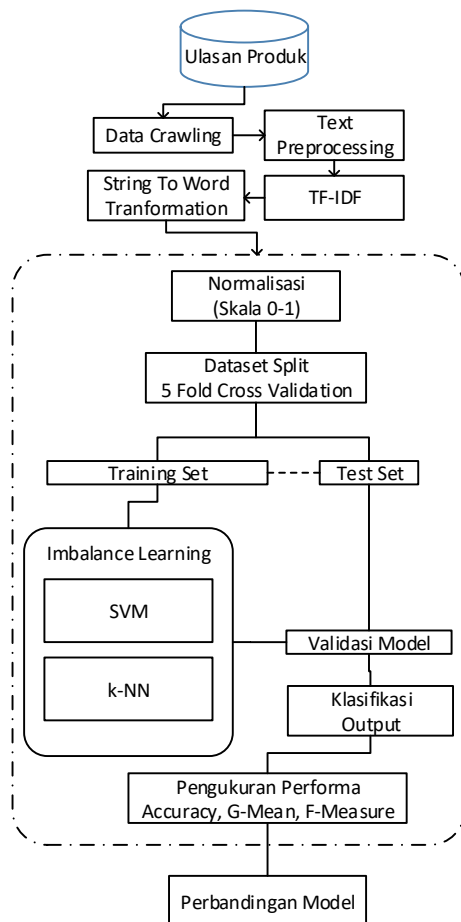
Gambar 2 memperlihatkan beberapa pola yang merupakan anggota dari dua buah kelas yaitu +1 dan -1. Warna merah merupakan representasi dari kelas -1, warna biru merupakan representasi dari kelas +1. Pekerjaan klasifikasi merupakan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut.

Garis putus-putus berfungsi sebagai alternatif garis pemisah antar kelas. *Hyperplane* pemisah terbaik antara kedua kelas dapat ditentukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya [2]. *Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing kelas. Pola yang paling dekat ini disebut sebagai *support vector* [9]. Garis tebal lurus pada gambar 2 menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk menemukan posisi *hyperplane* ini merupakan ide utama dari SVM.

3. METODOLOGI

3.1. Prosedur kerja

Prosedur kerja penelitian ditampilkan pada gambar 3. Data ulasan produk yang digunakan bersumber dari beberapa situs *e-commerce* yang ada di Indonesia.



Gambar 3. Prosedur kerja

3.2. Data penelitian

Penelitian ini menggunakan data ulasan produk yang dikumpulkan dari beberapa situs jual beli *online* di Indonesia. Pada gambar 2 berikut ini disajikan contoh ulasan produk *TrendyShoes* sepatu anak merek Denim.

Tabel 1 Sampel acak ulasan produk awal

Id	Ulasan Produk
27	Sepatunya bagus. tetapi ukuran nya kecil dari ukuran biasanya. Jd mesti pesan ukuran lebih besar dari ukuran biasanya
81	kiriman cepat sampai, &€ .BARANG SESUAI
89	Anak saya pas dipesan tidak mau..tapi ketika barang nya datang, dia suka banget.
39	beda sama yang digambar..!
120	warna kurang sesuai dengan gambar
16	pengirimannya lama banget &°Ä,ËœÄ¶
71	SANGAT SANGAT KECEWA !

3.3. Pemrosesan awal data

Setelah data ulasan produk berhasil di kumpulkan, maka tahap selanjutnya adalah tahap pemrosesan

awal (*pre-processing*) agar data ulasan produk dapat diterapkan pada algoritma *machine learning*. Tahapan *pre-processing* yang diterapkan adalah *Case Folding*, *Non Alpha Numeric Removal*, *Stop words Removal*, dan *Stemming*. Daftar *stop words* untuk Bahasa Indonesia terdiri atas 760 kata [10]. Algoritma *Stemming* yang diterapkan adalah algoritma *stemming* khusus untuk bahasa Indonesia yaitu algoritma nazief-Andriani [11].

Tabel 2 Hasil akhir *text pre-processing*

Id	Ulasan Produk
27	sepatu bagus ukur nya ukur jd mesti pesan ukur besar ukur
81	kirim cepat barang sesuai
89	anak pas pesan barang nya datang suka banget
39	beda gambar
120	warna kurang sesuai gambar
16	kirim banget
71	kecewa

3.4. TF-IDF

Untuk menentukan bobot setiap fitur pada data ulasan diterapkan algoritma TF-IDF. Hasil penerapan TF-IDF menghasilkan matriks data dengan dimensi 86 atribut x 1073 data. Dimensi data di atas masing sangat besar dan tidak efektif, sehingga atribut yang ada dievaluasi dan difilter.

3.5. Seleksi fitur data

Data ulasan produk terdiri atas 71 fitur, jumlah fitur tersebut masih sangat besar karena mengakibatkan dimensi data menjadi terlalu besar. Untuk membuang fitur fitur yang tidak signifikan terhadap proses text mining, dilakukan seleksi fitur menggunakan algoritma Coefficient feature selection (CFS) yang tersedia pada WEKA. Seleksi atribut menghasilkan sebanyak 14 fitur.

Pada tabel terdapat 14 kata kunci yang paling sering muncul di dalam ulasan produk, disusun berdasarkan abjad. Dari 14 kata kunci tersebut, terdapat 6 kata kunci positif dan 8 kata kunci negatif.

Tabel 3 Daftar kata kunci sentimen

No	keyword	Jenis Sentimen
1	awet	positif
2	bagus	positif
3	besar	negatif
4	cacat	negatif
5	cepat	positif
6	jelek	negatif
7	kecewa	negatif
8	lama	negatif

9	mengelupas	negatif
10	puas	positif
11	rusak	negatif
12	suka	positif
13	telat	negatif
14	terimakasih	positif

Pada tabel 4 ditampilkan sample 5 fitur atau kata kunci dengan bobot tertinggi TF-IDF. Kata kunci “Bagus” memiliki bobot dengan nilai TF-IDF terkecil yaitu 0,520.

Tabel 4 Daftar 5 fitur dengan bobot tertinggi data ulasan produk sepatu

No	Fitur	TF-IDF
1	Bagus	0,520
2	Sesuai	1,030
3	Terimakasih	1,236
4	Suka	1,464
5	Cepat	1,588

Penelitian ini menggunakan 3 jenis data ulasan produk. Data ulasan pertama terdiri atas 768 ulasan positif dan 564 ulasan negatif. Data ulasan kedua terdiri atas 768 ulasan positif dan 314 ulasan negatif. Data ulasan ketiga terdiri atas 268 ulasan positif dan 564 ulasan negatif.

Tabel 5 Deskripsi data ulasan

Data	#Pos	#Neg	#Ex	#IR	#Atts
Ulasan 1	768	564	1332	1.36	14
Ulasan 2	768	314	1082	2.44	14
Ulasan 3	268	564	832	2.10	14

Deskripsi lengkap data ulasan ditampilkan pada tabel 5. Pada tabel tersebut dideskripsikan banyak sentimen positif (#Pos.), banyak sentimen negatif (#Neg.), banyaknya data keseluruhan (#Ex.), tingkat ketidak seimbangan kelas (#IR.), dan banyaknya atribut (#Atts)

3.6. Teknik validasi dan evaluasi

Teknik evaluasi dan estimasi performa pada penelitian ini menggunakan skema *5-fold cross-validation*. Hal ini berarti, dataset ulasan dibagi menjadi 5 bagian atau *fold* yang sama, setiap *fold* berisi 20% dataset, kemudian dilakukan proses *learning* sebanyak 5 kali. Pada tabel ditampilkan hasil partisi data ulasan1. Teknik evaluasi dan pengukuran performa menerapkan *Area Under ROC Curve* (AUC). Pertimbangan penggunaan AUC karena AUC secara statistik lebih konsisten. AUC juga lebih baik dari metode akurasi (*accuracy*) dalam mengevaluasi perbandingan kinerja berbagai algoritma *classifier*

Tabel 6 Partisi data ulasan_1

Partisi Dataset	Jumlah data	Fungsi
ulasan1-1tra.dat	1065	Training
ulasan1-1tst.dat	267	Testing
ulasan1-2tra.dat	1065	Training
ulasan1-2tst.dat	267	Testing
ulasan1-3tra.dat	1065	Training
ulasan1-3tst.dat	267	Testing
ulasan1-4tra.dat	1065	Training
ulasan1-4tst.dat	267	Testing
ulasan1-5tra.dat	1065	Training
ulasan1-5tst.dat	267	Testing

3.7. Pengukuran Performa

Metode pengukuran performa memiliki peranan yang sangat penting untuk mengevaluasi kinerja suatu metode klasifikasi. *Confusion matrix* merupakan alat yang paling populer dalam mengevaluasi performa klasifikasi. Pada tabel berikut ditampilkan *confusion matrix* untuk kelas biner, sesuai dengan karakteristik data ulasan yaitu positif dan negatif

Tabel 7 Confution Matrix

Kelas	Prediktif Positif	Prediktif Negatif
Kelas Aktual Positif	TP	FN
Kelas Aktual Negatif	FP	TN

True Positive (TP) dan *True Negative* (TN) merupakan jumlah kelas positif dan negatif yang diklasifikasikan dengan tepat, *False Positive* (FP) dan *False Negative* (FN) merupakan jumlah kelas positif dan negatif yang tidak diklasifikasikan dengan tepat. Berdasarkan *confusion matrix* tersebut dapat ditentukan kriteria performa seperti *Accuracy*, *Precision*, *Recall*, *specificity*, *F-MEASURE*, *G-Mean* dan yang lainnya.

Akurasi (*Accuracy*) merupakan kriteria yang paling umum untuk mengukur kinerja klasifikasi, tetapi jika bekerja pada kelas tidak seimbang, kriteria ini kurang tepat karena kelas minoritas akan memiliki sumbangsih yang kecil pada kriteria *Accuracy*. Kriteria Penilaian yang disarankan adalah TP_{rate} , PP_{value} , *F-MEASURE* dan *G-Mean* [7]. *F-Measure* digunakan untuk mengukur klasifikasi kelas minoritas pada kelas tidak seimbang, dan indeks *G-mean* digunakan untuk mengukur performa keseluruhan (*overall classification*)

performance). Pada penelitian ini, performa klasifikasi menggunakan Accuracy, G-Mean, dan F-Measure.

$$Recall = TP_{rate} = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = PP_{value} = \frac{TP}{TP+FP} \quad (2)$$

$$Specificity = TN_{rate} = \frac{TN}{TN+FP} \quad (3)$$

$$G - Mean = \sqrt{TP_{rate} \cdot TN_{rate}} \quad (4)$$

4. Hasil Penelitian

Pada tabel 8 ditampilkan sampel acak hasil klasifikasi sentimen ulasan produk menggunakan SVM. Sentimen aktual adalah sentimen positif atau negatif yang sesungguhnya.

Tabel 8 Sampel acak klasifikasi ulasan produk

Id	Ulasan	Sentimen		Ket
		Aktual	Prediktif (SVM)	
701	ukuran tidak sesuai sepatu kecil ukur 35 tidak segitu besar kecewa berat...	Negatif	Positif	FP
62	barang bagus sesuai gambar terimakasih	Positif	Positif	TP
714	kualitas barang bagus kirim bagus bagus	Positif	Negatif	FN
603	anak suka sepatu alas kaki dalam tipis seprtinya tambah nih alas dalam lembut nyaman pakai over all keren cepat kirim tahan kualitas layanan terimakasih ☺☺	Positif	Negatif	FN
004	sepatu bagus sesuai gambar besaran pilih ukur	Positif	Positif	TP
015	Sepatu tidak sesuai gambar...!!! bohong harga mahal☹	Negatif	Negatif	TN

Dengan menggunakan skema validasi 5-fold cross validation, pada tabel 9 ditampilkan jumlah data training yang dikategorikan ke dalam TP, TN, FP, dan FN untuk data ulasan_1.

Tabel 9 Hasil validasi 5-fold CV data training ulasan_1

Fold	TP	TN	FP	FN
Fold-1	612	334	117	2
Fold-2	612	331	120	2

Fold-3	613	331	121	1
Fold-4	613	332	119	2
Fold-5	614	336	115	1
Rata-rata	612,8	332,8	118,4	1,6

Selanjutnya hasil rata-rata data training ada tabel 9 dimasukkan pada tabel 10 untuk menentukan rata-rata keseluruhan antara untuk data ulasan_1. Pada tabel 10, rata-rata True Positive (TP) pada data latih (training) sebanyak 612,5 data, rata-rata TP pada uji (testing) sebanyak 153,25 data. Selanjutnya rata-rata TP untuk data ulasan_1 adalah sebesar 382,87 data, rata-rata TN untuk data ulasan_1 adalah 208 data.

Tabel 10 Confusin matrix pengujian data ulasan_1

	TP	TN	FP	FN
Training	612,8	332,8	118,4	1,6
Testing	153,25	84	28,75	0,5
Rata-rata	382,88	208	74	1,125

Pada tabel 11, rata-rata True Positive (TP) sebanyak 382,88 data, rata-rata True Negative (TN) sebanyak 118,75 data, rata-rata False Positive (FP) sebanyak 38,25 data, dan rata-rata False Negative (FN) sebanyak 1,13 data.

Tabel 11 Confusin matrix pengujian data ulasan_2

	Training	Testing	Rata-rata
TP	612,5	153,25	382,88
TN	189,75	47,75	118,75
FP	61,5	15	38,25
FN	1,75	0,5	1,13

Pada tabel 12, rata-rata True Positive (TP) sebanyak 9,75 data, rata-rata True Negative (TN) sebanyak 279,13 data, rata-rata False Positive (FP) sebanyak 2,63 data, dan rata-rata False Negative (FN) sebanyak 35,25 data.

Tabel 12 Confusin matrix pengujian data ulasan_3

	TP	TN	FP	FN
Training	158	446,5	4,25	56,25
Testing	39,5	111,75	1	14,25
Rata-rata	98,75	279,13	2,63	35,25

Pada tabel 13, tabel 14, dan tabel 15 ditampilkan performa SVM pada klasifikasi sentimen pada data ulasan_1, ulasan_2, dan ulasan_3. Hasil SVM dibandingkan dengan metode K-Nearest Neighbor (K-NN). Dari ke tiga tabel tersebut hasil klasifikasi Sentimen menggunakan SVM lebih baik dari K-NN.

Tabel 13 Performa klasifikasi data ulasan_1

	SVM	KNN
Accuracy	88,83%	73,99%

G-Mean	85,91%	83,70%
F-Measure	91,15%	87,44%

Tabel 14 Performa klasifikasi data ulasan_2

	SVM	KNN
Accuracy	92,65%	78,66%
G-Mean	86,71%	83,23%
F-Measure	95,07%	83,09%

Tabel 15 Performa klasifikasi data ulasan_3

	SVM	KNN
Accuracy	90,80%	87,61%
G-Mean	85,30%	80,48%
F-Measure	83,71%	77,51%

5. Kesimpulan

Pada penelitian ini dilakukan analisis sentimen dan *text mining* pada ulasan produk. Data ulasan produk yang dikumpulkan kemudian dibagi menjadi tiga data yaitu ulasan_1, ulasan_2, dan ulasan_3. Ketiga data ulasan tersebut memiliki tingkat ketidak seimbangan kelas yang berbeda yaitu 1, 36; 2, 44 dan 2,10. Dengan menerapkan pengujian pada tiga jenis data sentimen yang berbeda diperoleh hasil bahwa *Support Vector Machine* dapat bekerja dengan baik pada data ulasan tidak seimbang. Performa *Support Vector Machine* dibandingkan dengan metode *K-Nearest Neighbor*. Hasil klasifikasi sentimen menggunakan *Support Vector Machine* lebih unggul dari *K-Nearest Neighbor*

reviews using machine learning techniques," *international Journal of Engineering and Technology*, vol. 7, no. 6, pp. 1–7, 2016.

- [6] R. Soni and K. J. Mathai, "Effective sentiment analysis of a launched product using clustering and decision trees," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 1, 2016.
- [7] H. Isah, P. Trundle, and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," in *2014 14th UK Workshop on Computational Intelligence (UKCI)*, 2014, pp. 1–7.
- [8] L. Kumar and P. K. Bhatia, "Text Mining: Concepts, Process and Applications," *Journal of Global Research in Computer Science*, vol. 4, no. 3, pp. 36–39, 2013.
- [9] I. C. R. Drajana, "Metode support vector machine dan forward selection prediksi pembayaran pembelian bahan baku kopra," vol. 9, p. 8, 2017.
- [10] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," *Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands*, 2003.
- [11] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, no. 4, pp. 1–33, 2007.
- [12] W. Prachuabsupakij and P. Doungpaisan, "Matching preprocessing methods for improving the prediction of student's graduation," in *Computer and Communications (ICCC), 2016 2nd IEEE International Conference on*, 2016, pp. 33–37.

References

- [1] Y. E. Ariska, W. Maharani, and M. S. Mubarak, "Peringkasan review produk berbasis fitur menggunakan semantic similarity scoring dan sentence clustering," p. 9.
- [2] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, p. 5, 2015.
- [3] P. Kowalchuk, "Implementing a Drilling Reporting Data Mining Tool Using Natural Language Processing Sentiment Analysis Techniques," in *SPE Middle East Oil and Gas Show and Conference*, 2019.
- [4] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," in *Computing, Communication & Automation (ICCCA), 2015 International Conference on*, 2015, pp. 933–937.
- [5] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie